(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) **EP 4 170 449 A1**

(12) **EUROPEAN PATENT APPLICATION**

(43) Date of publication:
26.04.2023 Bulletin 2023/17

(21) Application number: **22186467.1**

(22) Date of filing: **22.07.2022**

(51) International Patent Classification (IPC):
G05D 1/00 *(2006.01)* ***G05D 1/02*** *(2020.01)*
B25J 9/16 *(2006.01)* ***G01C 21/00*** *(2006.01)*
G06N 5/00 *(2023.01)* ***G06V 20/00*** *(2022.01)*
G10L 15/22 *(2006.01)*

(52) Cooperative Patent Classification (CPC):
G05D 1/0088; B25J 9/1689; G01C 21/00; G05D 1/0016; G05D 1/0248; G06N 5/00; G06V 20/00; G10L 15/22; G05D 2201/0211

(84) Designated Contracting States:
AL AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO RS SE SI SK SM TR
Designated Extension States:
BA ME
Designated Validation States:
KH MA MD TN

(30) Priority: **22.10.2021 IN 202121048303**

(71) Applicant: **Tata Consultancy Services Limited**
Maharashtra (IN)

(72) Inventors:
- **BANERJEE, SNEHASIS**
  700160 Kolkata - West Bengal (IN)
- **PURUSHOTHAMAN, BALAMURALIDHAR**
  560066 Bangalore - Karnataka (IN)
- **PRAMANICK, PRADIP**
  700160 Kolkata - West Bengal (IN)
- **SARKAR, CHAYAN**
  700160 Kolkata - West Bengal (IN)

(74) Representative: **Goddar, Heinz J.**
Boehmert & Boehmert
Anwaltspartnerschaft mbB
Pettenkoferstrasse 22
80336 München (DE)

(54) **SYSTEM AND METHOD FOR ONTOLOGY GUIDED INDOOR SCENE UNDERSTANDING FOR COGNITIVE ROBOTIC TASKS**

(57) Existing cognitive robotic applications follow a practice of building specific applications for specific use cases. However, the knowledge of the world and the semantics are common for a robot for multiple tasks. In this disclosure, to enable usage of knowledge across multiple scenarios, a method and system for ontology guided indoor scene understanding for cognitive robotic tasks is described where in scenes are processed based on techniques filtered based on querying ontology with relevant objects in perceived scene to generate a semantically rich scene graph. Herein, an initially manually created ontology is updated and refined in online fashion using external knowledge-base, human robot interaction and perceived information. This knowledge helps in semantic navigation, aids in speech, and text based human robot interactions. Further, in the process of performing the robotic tasks, the knowledgebase gets enriched, and the knowledge can be shared and used by other robots and services.

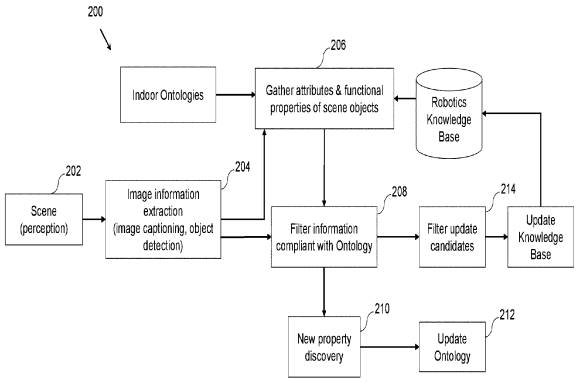

FIG. 2

EP 4 170 449 A1

Description

CROSS-REFERENCE TO RELATED APPLICATIONS AND PRIORITY

[0001] The present application claims priority from Indian patent application number 202121048303, filed on October 22, 2021.

TECHNICAL FIELD

[0002] The disclosure herein generally relates to the field of cognitive robotics and more specifically, to a system and method for an ontology guided indoor scene understanding for cognitive robotic tasks.

BACKGROUND OF THE INVENTION

[0003] The advent of low cost mobile robotic platform has seen a surge in usage of robots in our daily surroundings. The utility of a mobile robot has got expanded from personal usage to industries, shop floors, healthcare, and offices. Additionally, robots are equally used in collocated places and remote setups. The existing service robotic systems do not have a dedicated knowledgebase that can be updated and expanded using external knowledge as well as observations. Systems based on machine learning has an element of probability estimate errors and safety issues, whereas inclusion of commonsense and general knowledge can make decision making more semantically intelligent and reliable.

[0004] Ontology defines abstract concepts of a domain. The instances of the ontology form knowledgebase that ensures semantic interoperability of the knowledge among various applications (can be developed independently by different developers) of the domain. Considering the diverse range and ever evolving robotics applications, the ontologies need to be amended over time. Also, given the ontologies, the knowledgebase needs to be created by instantiating the classes and their properties as defined in the ontology for the environment where the robot is operating. This process is either done manually or automatically by collecting sensor-based data. With the change of the environment, the knowledgebase needs to be updated continuously, which should ensure consistency and non-conflict with the existing knowledge. It is important that when a robot interacts with a human being, the given knowledge needs to be grounded, i.e., both should have similar interpretation of the environment. Further, there is also problems in the existing cognitive robotics applications of building specific applications for specific use cases. However, the knowledge of the world and the semantics are common for a robot for multiple tasks.

SUMMARY OF THE INVENTION

[0005] Embodiments of the disclosure present technological improvements as solutions to one or more of the above-mentioned technical problems recognized by the inventors in conventional systems. For example, in one embodiment, a method and system for an ontology guided indoor scene understanding for cognitive robotic tasks is provided.

[0006] In one aspect, a processor-implemented method of an ontology guided indoor scene understanding for cognitive robotic tasks is provided. The method includes one or more steps such as obtaining, via an input/output interface, at least one navigation command for a robot from a user as an input, capturing one or more images of a scene by the robot, wherein each of the one or more images is a combination of Red Green Blue (RGB) image and its corresponding depth image, identifying one or more objects within the scene using a combination of image processing techniques, comprising an object detection, a captioning, and an ontology validation, querying an ontology to determine one or more properties related to each of the one or more identified objects, selecting at least one image processing technique from the combination of image processing techniques based on the determined one or more properties of each of the one or more objects and representation of the ontology to extract one or more attributes and relations from the scene, generating at least one scene graph using the extracted one or more attributes and relations to aid the robot in executing at least one navigation command, and finally updating a knowledge graph representation of a knowledge base based on the generated scene graph to facilitate an effective interaction between the user and the robot, wherein the knowledge graph representation represents grounded representation. It is to be noted that the knowledgebase comprises of a global ontology graph, an instance connected scene graph of a predefined environment represented using a multi-graph data structure, and individual scene graph for each of the one or more scenes. The knowledgebase herein is updated by one of exploration by the robot, a dialogue exchange of the user with the robot, and an external knowledge sources using one of filtering techniques to complement the seed ontology built specifically for the robotic domain and indoor applications. Furthermore, there is provision for technique linking of objects in scene with image processing techniques, in order to use only suitable scene processing technique based on declaration in ontology to a specific object, environment and task. There is further provision for technique linking to tasks, or in other words, the task instruction will be processed to extract keywords that can be mapped to tasks possible in that environment and enabling runtime actuation execution calls.

[0007] In another aspect, a system is configured for an ontology guided indoor scene understanding for cognitive robotic tasks is provided. The system includes an input/output interface configured to obtain at least one navigation command for a robot from a user as an input, one **or more hardware** processors and at least one

memory storing a plurality of instructions, wherein the one or more hardware processors are configured to execute the plurality of instructions stored in at least one memory.

[0008] Further, the system is configured to capture one or more images of a scene by the robot, wherein each of the one or more images is a combination of Red Green Blue (RGB) image and its corresponding depth image, identify one or more objects within the scene using a combination of image processing techniques, comprising an object detection, a captioning, and an ontology validation, querying an ontology to determine one or more properties related to each of the one or more identified objects. Further, the system is configured to select at least one image processing technique from the combination of image processing techniques based on the determined one or more properties of each of the one or more objects and representation of the ontology to extract one or more attributes and relations from the scene, generate at least one scene graph using the extracted one or more attributes and relations to aid the robot in executing at least one navigation command, and finally updates a knowledge graph representation of a knowledge base based on the generated scene graph to facilitate an effective interaction between the user and the robot, wherein the knowledge graph representation represents grounded representation. It is to be noted that the knowledgebase comprises of a global ontology graph, an instance connected scene graph of a predefined environment represented using a multi-graph data structure, and individual scene graph for each of the one or more scenes. The knowledgebase herein is updated by one of exploration by the robot, a dialogue exchange of the user with the robot, and an external knowledge sources using one of filtering techniques to complement the seed ontology built specifically for the robotic domain and indoor applications. Furthermore, there is provision for technique linking of objects in scene with image processing techniques, in order to use only suitable scene processing technique based on declaration in ontology to a specific object, environment and task. There is further provision for technique linking to tasks, or in other words, the task instruction will be processed to extract keywords that can be mapped to tasks possible in that environment and enabling runtime actuation execution calls.

[0009] In yet another aspect, one or more non-transitory machine-readable information storage mediums are provided comprising one or more instructions, which when executed by one or more hardware processors causes a method of an ontology guided indoor scene understanding for cognitive robotic tasks is provided. The method includes one or more steps such as obtaining, via an input/output interface, at least one navigation command for a robot from a user as an input, capturing one or more images of a scene by the robot, wherein each of the one or more images is a combination of Red Green Blue (RGB) image and its corresponding depth image, identifying one or more objects within the scene using a combination of image processing techniques, comprising an object detection, a captioning, and an ontology validation, querying an ontology to determine one or more properties related to each of the one or more identified objects, selecting at least one image processing technique from the combination of image processing techniques based on the determined one or more properties of each of the one or more objects and representation of the ontology to extract one or more attributes and relations from the scene, generating at least one scene graph using the extracted one or more attributes and relations to aid the robot in executing at least one navigation command, and finally updating a knowledge graph representation of a knowledge base based on the generated scene graph to facilitate an effective interaction between the user and the robot, wherein the knowledge graph representation represents grounded representation. It is to be noted that the knowledgebase comprises of a global ontology graph, an instance connected scene graph of a predefined environment represented using a multi-graph data structure, and individual scene graph for each of the one or more scenes. The knowledgebase herein is updated by one of exploration by the robot, a dialogue exchange of the user with the robot, and an external knowledge sources using one of filtering techniques to complement the seed ontology built specifically for the robotic domain and indoor applications. Furthermore, there is provision for technique linking of objects in scene with image processing techniques, in order to use only suitable scene processing technique based on declaration in ontology to a specific object, environment and task. As an example of technique linking, in the ontology **for an object 'Television', the decision-making** module will query the ontology to get the relevant properties **like** *'hasColor'* **and** *"hasShape'* **and fetch the** corresponding technique link in to execute the image processing technique to detect the color and shape in the designated region in the image scene at runtime. There is further provision for technique linking to tasks, or in other words, the task instruction will be processed to extract keywords that can be mapped to tasks possible in that environment and enabling runtime actuation execution calls.

[0010] It is to be understood that the foregoing general descriptions and the following detailed description are exemplary and explanatory only and are not restrictive of the invention, as claimed.

BRIEF DESCRIPTION OF THE DRAWINGS

[0011] The accompanying drawings, which are incorporated in and constitute a part of this disclosure, illustrate exemplary embodiments and, together with the description, serve to explain the disclosed principles:

FIG. 1 illustrates a network diagram of an exemplary system for an ontology guided indoor scene understanding for cognitive robotic tasks in accordance

with some embodiments of the present disclosure.

FIG. 2 illustrates a functional block diagram to illustrate the exemplary system in accordance with some embodiments of the present disclosure.

FIG. 3 is a schematic diagram to illustrating an example of the robot's decision making based on the knowledge base in accordance with some embodiments of the present disclosure.

FIG. 4 illustrates a functional block diagram to illustrate the ontology extension and population in accordance with some embodiments of the present disclosure.

FIG. 5 is a flow diagram to illustrate a method of ontology guided indoor scene understanding for cognitive robotic tasks in accordance with some embodiments of the present disclosure.

[0012] It should be appreciated by those skilled in the art that any block diagrams herein represent conceptual views of illustrative systems and devices embodying the principles of the present subject matter. Similarly, it will be appreciated that any flow charts, flow diagrams, and the like represent various processes, which may be substantially represented in computer readable medium and so executed by a computer or processor, whether or not such computer or processor is explicitly shown.

DETAILED DESCRIPTION OF EMBODIMENTS

[0013] Exemplary embodiments are described with reference to the accompanying drawings. In the figures, the left-most digit(s) of a reference number identifies the figure in which the reference number first appears. Wherever convenient, the same reference numbers are used throughout the drawings to refer to the same or like parts. While examples and features of disclosed principles are described herein, modifications, adaptations, and other implementations are possible without departing from the scope of the disclosed embodiments.

[0014] The embodiments herein provide a method and system for ontology guided indoor scene understanding for cognitive robotic tasks. Herein, an ontology is updated online using external knowledge-base and observed information. This knowledge helps in semantic navigation, aids in speech, and text based human robot interactions. Further, in the process of performing the robotic tasks, the knowledgebase gets enriched, and the knowledge can be shared and used by other robots.

[0015] It is to be noted that when domain is limited like an indoor environment with limited number of possible objects, the decision making becomes much better equipped if the knowledge of the environment is grounded in an ontology that can be referenced for generation of semantic summary of scenes via scene graphs. This leads to reliable scene graphs that can be linked to external knowledge sources. Further, the scene graph generation is an ontology driven, hence chance of semantic errors at knowledge level is very low and processing also happens fast as only selective image processing techniques are run as per possible predicates and range of predicate values neglect erroneous output from image processing techniques.

[0016] Ontologies define the abstract concepts of a domain. The instances of the ontology form a knowledgebase that ensures semantic interoperability of the knowledge among various applications (can be developed independently by different developers) of the domain. Thus, the IEEE Standard Association's Robotics Society has formed Ontologies for Robotics and Automation (ORA) Working Group. Considering the diverse range and ever evolving robotics applications, the ontologies need to be amended over time. Also, given the ontologies, the knowledgebase needs to be created by instantiating the classes and their properties as defined in the ontology for the environment where the robot is operating. This process is either done manually or automatically by collecting sensor-based data. With the change of the environment, the knowledgebase needs to be updated continuously, which should ensure consistency and non-conflict with the existing knowledge.

[0017] Further, when a robot interacts with a human being, the given knowledge needs to be grounded, i.e., both should have similar interpretation of the environment. As the scope of human-robot interaction is increasing, this is becoming an active research domain to define ontologies and create knowledgebase that can facilitate effective interaction between a human and robot. Moreover, a lot of knowledge can be populated from the dialogue exchange with a human (explicit **or implicit). There are** some **techniques to populate knowledge based on robot's** sensor data, but no such work that enables people of grounded knowledge from spatial dialogue with a human-robot setting.

[0018] It would be appreciated that herein a two-way communication in terms of ontology access is used. The ontology is used to understand scene as well as spatial dialogue keyword references in order to carry out cognitive robotic tasks. Apart from this, the ontology and instance of scene graph is also refined by robotic exploration of the environment to learn property instances of objects, as well as direct linking through dialogue exchange with robot to know more about the world model.

[0019] It is to be noted that the prevalent approaches of knowledge base are to use some in memory knowledge without a formal definition of the knowledge representation. This poses a challenge in near future when these solutions need to be integrated to form a bigger solution. Thus, a standard way to represent the knowledge would ensure interoperability of various smaller solutions. Moreover, the scope of interacting with human (operator, coworker) are also increasing as there is growing trend of mix workforce. As a result, there should be a common understanding of knowledge not among various robotic sub-solutions, but between a human and robotic agent as well. Thus, knowledge representation should facilitate grounded representation. Now, any ro-

botic environment is dynamic, and some information needs continuous update. Also, it may not be feasible to populate the entire knowledgebase before the deployment. Thus, there is a need to continuously update the knowledgebase with new/modified information.

[0020] Referring now to the drawings, and more particularly to FIG. 1 through FIG. 5, where similar reference characters denote corresponding features consistently throughout the figures, there are shown preferred embodiments and these embodiments are described in the context of the following exemplary system and/or method.

[0021] FIG. 1 illustrates a block diagram of a system (100) for an ontology guided indoor scene understanding for cognitive robotic tasks, in accordance with an example embodiment. Although the present disclosure is explained considering that the system (100) is implemented on a server, it may be understood that the system (100) may comprise one or more computing devices (102), such as a laptop computer, a desktop computer, a notebook, a workstation, a cloud-based computing environment and the like. It will be understood that the system (100) may be accessed through one or more input/output interfaces 104-1, 104-2... 104-N, collectively referred to as I/O interface (104). Examples of the I/O interface (104) may include, but are not limited to, a user interface, a portable computer, a personal digital assistant, a handheld device, a smartphone, a tablet computer, a workstation, and the like. The I/O interface (104) are communicatively coupled to the system (100) through a network (106).

[0022] In an embodiment, the network (106) may be a wireless or a wired network, or a combination thereof. In an example, the network (106) can be implemented as a computer network, as one of the different types of networks, such as virtual private network (VPN), intranet, local area network (LAN), wide area network (WAN), the internet, and such. The network (106) may either be a dedicated network or a shared network, which represents an association of the different types of networks that use a variety of protocols, for example, Hypertext Transfer Protocol (HTTP), Transmission Control Protocol/Internet Protocol (TCP/IP), and Wireless Application Protocol (WAP), to communicate with each other. Further, the network (106) may include a variety of network devices, including routers, bridges, servers, computing devices, storage devices. The network devices within the network (106) may interact with the system (100) through communication links.

[0023] The system (100) supports various connectivity options such as BLUETOOTH®, USB, ZigBee, and other cellular services. The network environment enables connection of various components of the system (100) using any communication link including Internet, WAN, MAN, and so on. In an exemplary embodiment, the system (100) is implemented to operate as a standalone device. In another embodiment, the system (100) may be implemented to work as a loosely coupled device to a smart computing environment. Further, the system (100) comprises at least one memory with a plurality of instructions, one or more databases (110), and one or more hardware processors (108) which are communicatively coupled with the at least one memory to execute a plurality of modules therein.

[0024] The one or more I/O interfaces (104) are configured to obtain at least one navigation command for a robot from a user as an input. The one or more I/O interfaces (104) are also configured to enable the user to update a knowledge graph representation of a knowledge base based on the generated scene graph to facilitate an effective interaction between the user and the robot The components and functionalities of the system (100) are described further in detail.

[0025] Referring FIG. 2, a functional block diagram (200) to illustrate the system (100), wherein the system (100) is configured to capture one or more Red Green Blue - Depth (RGB-D) images of a scene by the robot. Each of the RGB-D image is a combination of RGB image and its corresponding depth image. The depth image is an image channel in which each pixel relates to a distance between an image plane and a corresponding object in the RGB image. Further, the system identifies one or more objects within the scene using a combination of image processing techniques, comprising an object detection, a captioning, and an ontology validation (202).

[0026] Further, the system (100) is configured to query an ontology to determine one or more properties related to each of the one or more identified objects (202). It is to be noted the system (100) creates a robotic base ontology that align with existing ontologies. Then via a knowledge service, the corresponding requests for respective robotic tasks may be served. In turn, via perception processing, the knowledge may also get enhanced.

[0027] wherein for knowledge guided scene graph generation, the system (100) describe deep learning based image processing techniques to detect object and certain properties about the objects (206). To be able to interact with a human being (in natural language), a robot should be able to map human task instruction to its capability and referred objects to the environment. This process of grounding is enabled by a knowledge guided scene graph. The image processing techniques would provide a certain set of properties irrespective of the environment or the human agent. The knowledge guided (grounded) scene graph generation helps to bridge the gap that requires costly customization of the image processing techniques yet provide an easily customizable scene graph.

[0028] It is to be noted that the robot needs to understand where in an environment it is in and for input it has sensors like camera, depth sensor, microphone, infrared sensor, etc. Based on this sensing inputs, the robot may generate a scene graph, a graph containing objects in a scene and corresponding relationships of the objects. Traditional scene graphs have just nodes in the graph

with some unnamed relations to it. However, in this case, because there is the ontology from where the objects detected by an object detection technique can be found and check for properties, the resultant scene graph contains relation edges that comes from the ontology validation itself (208).

[0029] Referring FIG. 3, a schematic diagram (300) to illustrate usage of knowledge semantics in dialogue and semantic navigation in one instruction in accordance with some embodiments of the present disclosure. **Wherein if a** *'cup'* is detected as an object in an image scene processing, then the system (100) is configured to search ontology to find what relations does cup have with the scene. It sees that in ontology relations are size, color, on top of, near. Using image processing technique, the relevant mapping can be done. So accordingly based on whether actual relation exists in the ontology or not, the relations in scene can be searched and populated. Also, if using an image caption technique to generate caption relations for the images, there are usually wrong caption generations that are encountered. Hence using the ontology to check if the captions are correct or not can be a way out to make the scene graph logically consistent. As an example, **if a caption generator says that** *"table is on top of cup",* **using ontology the reverse** correct relation can be established by checking the direction of the edge between nodes. In the ontology, the relations can be directional (on top of) or bi-directional (near/ proximity). This matching can be done by using a look up method whenever objects are found in the scene.

[0030] Further, the system (100) selects at least one image processing technique from the combination of image processing techniques based on the determined one or more properties of each of the one or more objects and representation of the ontology to extract one or more attributes and relations from the scene. It would be appreciated that given a generic image processing technique, the outcome of it may need to be processed further for particular application or requirement. This filtering requires a set of techniques for the process, which technique is applicable for a particular scenario is guided by the ontological representation (208).

[0031] There is provision for technique linking of objects in scene with image processing techniques, in order to use only suitable scene processing techniques based on declaration in ontology to a specific object, environment and task. There is further provision for technique linking to tasks, or in other words, the task instruction may be processed to extract keywords that can be mapped to tasks possible in that environment and enabling runtime actuation execution calls. **As an example of technique linking, in the ontology for an object 'Television', the** decision-making module will query the ontology to get the relevant properties like *'hasColor'* **and** *"hasShape'* **and fetch the corresponding technique link in to** execute the image processing technique to detect the color and shape in the designated region in the image scene at runtime.

[0032] In another example, wherein **for the** *"red cup"* **instruction, when the object is detected as** *'cup',* **from ontology lookup, the object properties relevant to the object** *'cup'* **is fetched. Color is one such property mentioned in ontology.** Further, the system is configured to invoke a technique to detect color, that uses image pixel processing. This invocation is done by a path stored in the ontology, **so that when** *'color'* **property is activated, the corresponding technique mapped to** it is called and the result is checked within the range of colors.

[0033] In this way, instead of calling each type of feature technique for each and every object, only the relevant ones are called based on filtering on two conditions (210): a) whether the object has features (edge connected nodes) that are relevant for it, b) the feature set relevant for the given instruction. As an **example, to detect** *'red cup'* **as per instruction as soon the keywords are processed,** *'red'* **is marked as color,** *'cup'* **is an indoor object. Then, the** technique to detect '*red*' **feature is called from** technique mapping stored in the ontology and at **runtime processing happens to give output. If cup found in scene is** *'blue',* **then it can be inferred that this is a** *'blue'* **cup. Similarly, for another object** '*TV*', **the color** is always black when off and random when on, a shape detector technique to output *'rectangle within rectangle'* **can be called to check if it is indeed a** '*TV*' instead of a picture frame. Similarly based on color, '*TV*' and window can be disambiguated.

[0034] In another embodiment, the system (100) is configured to generate at least one scene graph using the extracted one or more attributes and relations to aid the robot in executing at least one navigation command. The generated at least one scene graph is used to update a knowledge graph representation of a knowledge base to facilitate an effective interaction between the user and the robot (214).

[0035] In another aspect, wherein the user is stating facts about the environment explicitly or implicitly during other conversation is a good source of knowledge. The system (100) extracts knowledge from the dialogue with the user, filter it according to the knowledge representation structure, and then store it in the knowledgebase. The keywords are checked with ontology for possible existing entries in the knowledgebase itself, and new information is added based on a confidence estimation technique.

[0036] In another example, wherein **the information that** *'cup is mine'* may **result in knowledge that** '*cup*' **as an object belongs to owner** *'user X',* **as each object** of personal type will have an owner. This information can be learnt via a dialogue. Also, in some cases, an object may have a property instance that is not getting detected or is a new property altogether. That can be added based on user confirmation **- like** *'cup'* **is of** *'gray'* **color, or** *'cup'* **has property of** '*handle*' **like** a mug, this information was not there earlier in the ontology. This happens by using finding nodes matching the keywords and creating new

nodes of property values in case of **specific indoor scene; or new edge properties like** '*cup*' has component '*handle*'.

[0037] It would be appreciated that the knowledge graph representation facilitates grounded representation. The knowledgebase comprises of a global ontology graph, an instance connected scene graph of a predefined environment represented using a multi-graph data structure, and individual scene graph for each of the one or more scenes (216). Further, the knowledgebase is updated by one of exploration by the robot, a dialogue exchange of the user with the robot, and an external knowledge sources using one of filtering techniques to complement the seed ontology built specifically for the robotic domain and indoor applications.

[0038] In another example, wherein if it is found that an object lies in an odd location for that particular environment, then the same is updated as a special **instance case. Like,** '*oven*' **is lying in** *'living room'* **instead of kitchen. This special** instance can be updated in the scene graph of that environment if that is the practice and later it can be used accordingly. Also, if there is a general property found, that is also added if the frequency of occurrence is generic enough in multiple **environments. Like,** '*glass*' **can contain** *'water'* **after disambiguation of glass as** an object (not an element) and if water is detected around glass in multiple scenes, then by inferencing (reasoning) to find the cause, this relation can be established, which was not there in the ontology beforehand.

[0039] Moreover, a global knowledgebase which can be shared across multiple robots and software services as a query service. The global knowledge graph is used for cognitive robotics tasks such as and not limited to manipulation, navigation, dialogue exchange; and the global knowledge graph is updated based on inputs derived from feedback of scene exploration by robot, manipulation of objects and new knowledge mining using dialogue exchange and optionally external knowledge sources in graph compliant format.

[0040] Referring FIG. 4, a functional block diagram (400) to illustrate ontology extension and population in accordance with some embodiments of the present disclosure. Initially, a seed ontology is created manually by declaring concepts relevant to the robotic domain and the indoor environment that comprises of commonly occurring objects and their object properties (like color) along with inter object properties (like proximity relation) (402). This ontology is aligned with a standard Suggested Upper Merged Ontology (SUMO) to make it semantic web coherence compliant and also with an abstract Core Ontology for Robotics and Automation (CORA) to make it compliant with acceptable robotic standardization. When a task instruction is given to the robot (404), the robot used perception information to make sense of where it is and what is sensed (like visible in camera sensor) (406) and then based on task instruction type and current scene analysis requests the knowledgebase to get relevant information in form a service call (408). The ontology can also be updated or refined as a service if perception processing entails new generic information not earlier there in the ontology (410).

[0041] In another embodiment, a granular knowledgebase is developed for the robot that complies to the known standards of robotics ontology definition. This in turn ensures that the knowledge stored using this representation can be usable by any application who understanding the definition of the representation. The knowledge representation enables easy update of the existing data and it supports grounded representation of the knowledge with semantically rich information. This ensures usability of it in human-robot interaction scenario. The knowledge format is multi-graph, which is generic to support subject-predicate-object relations in form of node-edge-node. This specific choice of format allows multiple edge instances between two nodes as well as edge properties. This is useful if the user wishes to map it to a Planning Domain Definition Language (PDDL), well graph network analysis, and forming queries on SPARQL. The graph has options of **direction embedded in the property as an example** *'proximity'* **is a bi-directional property whereas the** *'hasColor'* **is a unidirectional property.**

[0042] Referring FIG. 5, to illustrate a flowchart (500), for an ontology guided indoor scene understanding for cognitive robotic tasks, in accordance with an example embodiment. Initially, at the step (502), obtaining, via an input/output interface, at least one navigation command for a robot from a user as an input.

[0043] At the next step (504), capturing one or more images of a scene by the robot, wherein each of the one or more images is a combination of Red Green Blue (RGB) image and its corresponding depth image.

[0044] At the next step (506), identifying one or more objects within the scene using a combination of image processing techniques, comprising an object detection, a captioning, and an ontology validation.

[0045] At the next step (508), querying an ontology to determine one or more properties related to each of the one or more identified objects.

[0046] At the next step (510), selecting at least one image processing technique from the combination of image processing techniques based on the determined one or more properties of each of the one or more objects and representation of the ontology to extract one or more attributes and relations from the scene.

[0047] At the next step (512), generating at least one scene graph using the extracted one or more attributes and relations to aid the robot in executing at least one navigation command.

[0048] At the last step (514), updating a knowledge graph representation of a knowledge base based on the generated scene graph to facilitate an effective interaction between the user and the robot, wherein the knowledge graph representation facilitates grounded representation.

[0049] The written description describes the subject matter herein to enable any person skilled in the art to make and use the embodiments. The scope of the subject matter embodiments is defined by the claims and may include other modifications that occur to those skilled in the art. Such other modifications are intended to be within the scope of the claims if they have similar elements that do not differ from the literal language of the claims or if they include equivalent elements with insubstantial differences from the literal language of the claims.

[0050] The embodiments of present disclosure herein address unresolved problem in cognitive robotics applications of building specific applications for specific use cases. However, the knowledge of the world and the semantics are common for a robot for multiple tasks. In this disclosure, to enable usage of knowledge across multiple scenarios, a method and system for ontology guided indoor scene understanding for cognitive robotic tasks is described where in scenes are processed based on techniques filtered based on querying ontology with relevant objects in perceived scene to generate a semantically rich scene graph. Herein, an initially manually created ontology is updated and refined in online fashion using external knowledge-base, human robot interaction and observed information. This knowledge helps in semantic navigation, aids in speech, and text based human robot interactions. Further, in the process of performing the robotic tasks, the knowledgebase gets enriched and the knowledge can be shared and used by other robots.

[0051] It is to be understood that the scope of the protection is extended to such a program and in addition to a computer-readable means having a message therein; such computer-readable storage means contain program-code means for implementation of one or more steps of the method, when the program runs on a server or mobile device or any suitable programmable device. The hardware device can be any kind of device which can be programmed including e.g., any kind of computer like a server or a personal computer, or the like, or any combination thereof. The device may also include means which could be e.g., hardware means like e.g., an application-specific integrated circuit (ASIC), a field-programmable gate array (FPGA), or a combination of hardware and software means, e.g., an ASIC and an FPGA, or at least one microprocessor and at least one memory with software modules located therein. Thus, the means can include both hardware means, and software means. The method embodiments described herein could be implemented in hardware and software. The device may also include software means. Alternatively, the embodiments may be implemented on different hardware devices, e.g., using a plurality of CPUs.

[0052] The embodiments herein can comprise hardware and software elements. The embodiments that are implemented in software include but are not limited to, firmware, resident software, microcode, etc. The functions performed by various modules described herein may be implemented in other modules or combinations of other modules. For the purposes of this description, a computerusable or computer readable medium can be any apparatus that can comprise, store, communicate, propagate, or transport the program for use by or in connection with the instruction execution system, apparatus, or device.

[0053] The illustrated steps are set out to explain the exemplary embodiments shown, and it should be anticipated that ongoing technological development will change the manner in which particular functions are performed. These examples are presented herein for purposes of illustration, and not limitation. Further, the boundaries of the functional building blocks have been arbitrarily defined herein for the convenience of the description. Alternative boundaries can be defined so long as the specified functions and relationships thereof are appropriately performed. Alternatives (including equivalents, extensions, variations, deviations, etc., of those described herein) will be apparent to persons skilled in the relevant art(s) based on the teachings contained herein. Such alternatives fall within the scope of the disclosed embodiments. Also, the words **"comprising," "having," "containing," and "including," and other similar forms are** intended to be equivalent in meaning and be open ended in that an item or items following any one of these words is not meant to be an exhaustive listing of such item or items or meant to be limited to only the listed item or items. It must also be **noted that as used herein and in the appended claims, the singular forms "a," "an," and "the" include plural references unless the context clearly dictates otherwise.**

[0054] Furthermore, one or more computer-readable storage media may be utilized in implementing embodiments consistent with the present disclosure. A computer-readable storage medium refers to any type of physical memory on which information or data readable by a processor may be stored. Thus, a computer-readable storage medium may store instructions for execution by one or more processors, including instructions for causing the processor(s) to perform steps or stages consistent with the embodiments described herein. The term **"computer-readable medium" should be understood to include tangible items and** exclude carrier waves and transient signals, i.e., be non-transitory. Examples include random access memory (RAM), read-only memory (ROM), volatile memory, nonvolatile memory, hard drives, CD ROMs, DVDs, flash drives, disks, and any other known physical storage media.

[0055] It is intended that the disclosure and examples be considered as exemplary only, with a true scope of disclosed embodiments being indicated by the following claims.

Claims

1. A processor-implemented method (500) comprising steps of:

obtaining (502), via an input/output interface (104), at least one navigation command for a robot from a user as an input;

capturing (504), via one or more hardware processors (108), one or more images of a scene by the robot, wherein each of the one or more images is a combination of Red Green Blue (RGB) images and their corresponding depth images;

identifying (506), via the one or more hardware processors (108), one or more objects within the scene using a combination of image processing techniques, wherein the image processing techniques comprising an object detection technique, a captioning technique, and an ontology validation technique;

querying (508), via the one or more hardware processors (108), an ontology to determine one or more properties related to each of the one or more identified objects;

selecting (510), via the one or more hardware processors (108), at least one image processing technique from the combination of image processing techniques based on the determined one or more properties of each of the one or more objects and representation of the ontology to extract one or more attributes and relations from the scene;

generating (512), via the one or more hardware processors (108), at least one scene graph using the extracted one or more attributes and relations to aid the robot in executing at least one navigation command; and

updating (514), via the one or more hardware processors (108), a knowledge graph representation of a knowledge base based on the generated scene graph to facilitate an effective interaction between the user and the robot, wherein the knowledge graph representation represents grounded representation.

2. The processor-implemented method (500) of claim 1, further comprising a global knowledgebase shared across a plurality of robots and software services as a query service.

3. The processor-implemented method (500) of claim 1, wherein the depth image is an image channel in which each pixel relates to a distance between an image plane and a corresponding object in the RGB image.

4. The processor-implemented method (500) of claim 1, wherein the knowledgebase comprises a global ontology graph, an instance connected scene graph of a predefined environment represented using a multi-graph data structure, and an individual scene graph for each of the one or more scenes.

5. The processor-implemented method (500) of claim 1, wherein the knowledgebase is updated by one of exploration by the robot, a dialogue exchange of the user with the robot, and an external knowledge source using one of filtering techniques to complement a seed ontology built specifically for the robotic domain and indoor applications.

6. The processor-implemented method (500) of claim 1, wherein the scene graph generation using image processing techniques is validated by object-property relations and range of values declared in the ontology.

7. The processor-implemented method (500) of claim 1, wherein the global knowledge graph is used for cognitive robotics tasks, wherein the cognitive robotics tasks include manipulation, navigation, and dialogue exchange.

8. The processor-implemented method (500) of claim 1, wherein the knowledge graph representation is updated based on inputs derived from feedback of scene exploration by the robot, manipulation of one or more objects and a new knowledge mining using the dialogue exchange and optionally external knowledge sources in graph compliant format.

9. A system (100) comprising:

an **input/output interface (104) to obtain** at least one navigation command for a robot from a user as an input;

one **or more hardware processors (108);**

a memory in communication with the one or more hardware processors (108), wherein the one or more hardware processors (108) are configured to execute programmed instructions stored in the memory, to:

capture one or more images of a scene by the robot, wherein each of the one or more images is a combination of Red Green Blue (RGB) images and their corresponding depth image;

identify one or more objects within the scene using a combination of image processing techniques, wherein the image processing techniques comprising an object detection technique, a captioning technique, and an ontology validation technique;

query an ontology to determine one or more properties related to each of the one or more identified objects;

select at least one image processing technique from the combination of image processing techniques based on the determined one or more properties of each of the

one or more objects and representation of the ontology to extract one or more attributes and relations from the scene;
generate at least one scene graph using the extracted one or more attributes and relations to aid the robot in executing at least one navigation command; and
update a knowledge graph representation of a knowledge base based on the generated scene graph to facilitate an effective interaction between the user and the robot, wherein the knowledge graph representation represents grounded representation.

10. The system (100) of claim 9, further comprising a global knowledgebase shared across a plurality of robots and software services as a query service.

11. The system (100) of claim 9, wherein the depth image is an image channel in which each pixel relates to a distance between an image plane and a corresponding object in the RGB image.

12. The system (100) of claim 9, wherein the knowledgebase comprises a global ontology graph, an instance connected scene graph of a predefined environment represented using a multi-graph data structure, and an individual scene graph for each of the one or more scenes.

13. The system (100) of claim 9, wherein the knowledgebase is updated by one of exploration by the robot, a dialogue exchange of the user with the robot, and an external knowledge source using one of filtering techniques to complement a seed ontology built specifically for the robotic domain and indoor applications.

14. The system (100) of claim 9, wherein the scene graph generation using image processing techniques is validated by object-property relations and range of values declared in the ontology.

15. A non-transitory computer readable medium storing one or more instructions which when executed by one or more processors on a system, cause the one or more processors to perform method comprising:

obtaining (502), via an input/output interface, at least one navigation command for a robot from a user as an input;
capturing (504), via one or more hardware processors (108), one or more images of a scene by the robot, wherein each of the one or more images is a combination of Red Green Blue (RGB) image and its corresponding depth image;
identifying (506), via the one or more hardware processors (108), one or more objects within the scene using a combination of image processing techniques, comprising an object detection, a captioning, and an ontology validation;
querying (508), via the one or more hardware processors (108), an ontology to determine one or more properties related to each of the one or more identified objects;
selecting (510), via the one or more hardware processors (108), at least one image processing technique from the combination of image processing techniques based on the determined one or more properties of each of the one or more objects and representation of the ontology to extract one or more attributes and relations from the scene;
generating (512), via the one or more hardware processors (108), at least one scene graph using the extracted one or more attributes and relations to aid the robot in executing at least one navigation command; and
updating (514), via the one or more hardware processors (108), a knowledge graph representation of a knowledge base based on the generated scene graph to facilitate an effective interaction between the user and the robot, wherein the knowledge graph representation represents grounded representation.

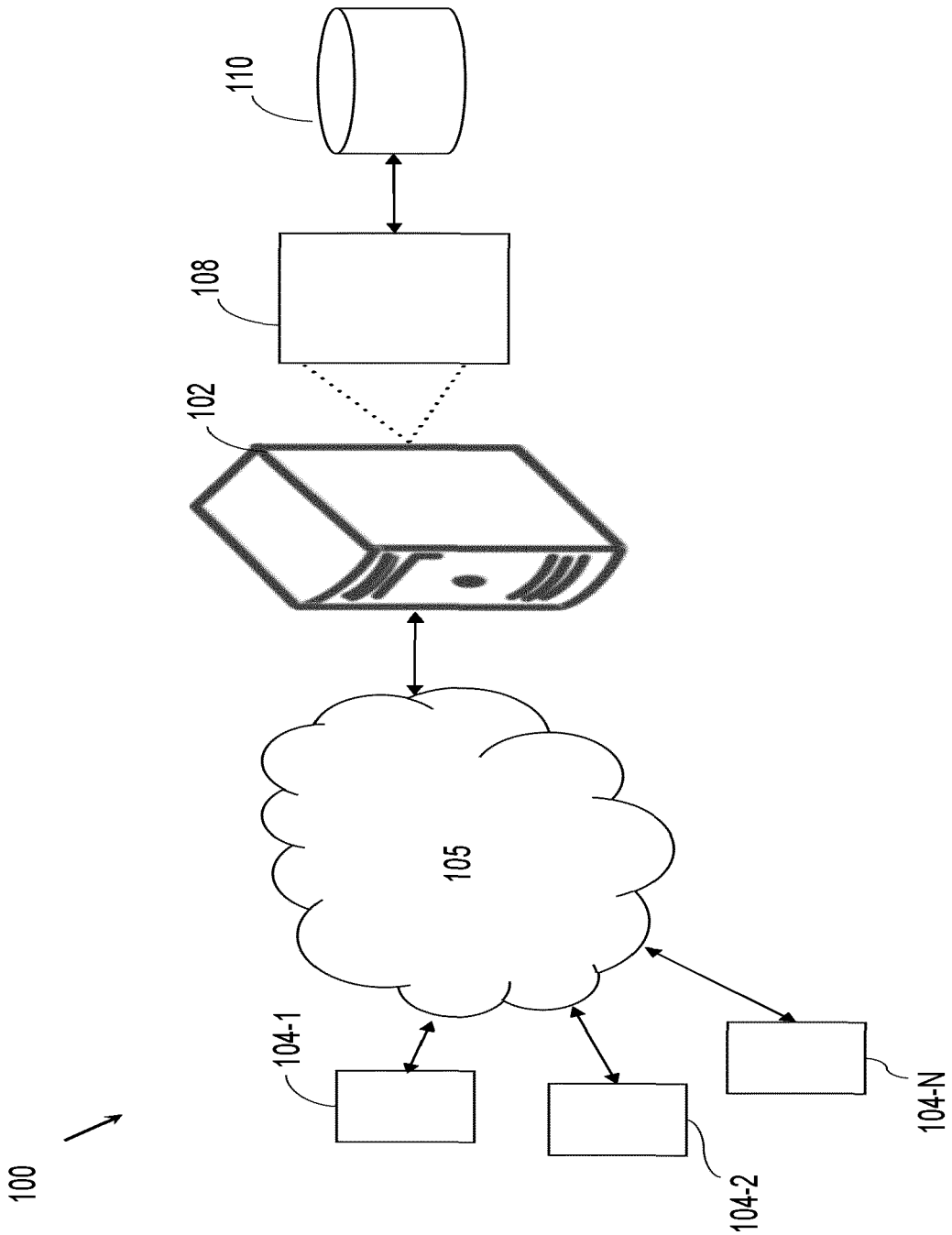

FIG. 1

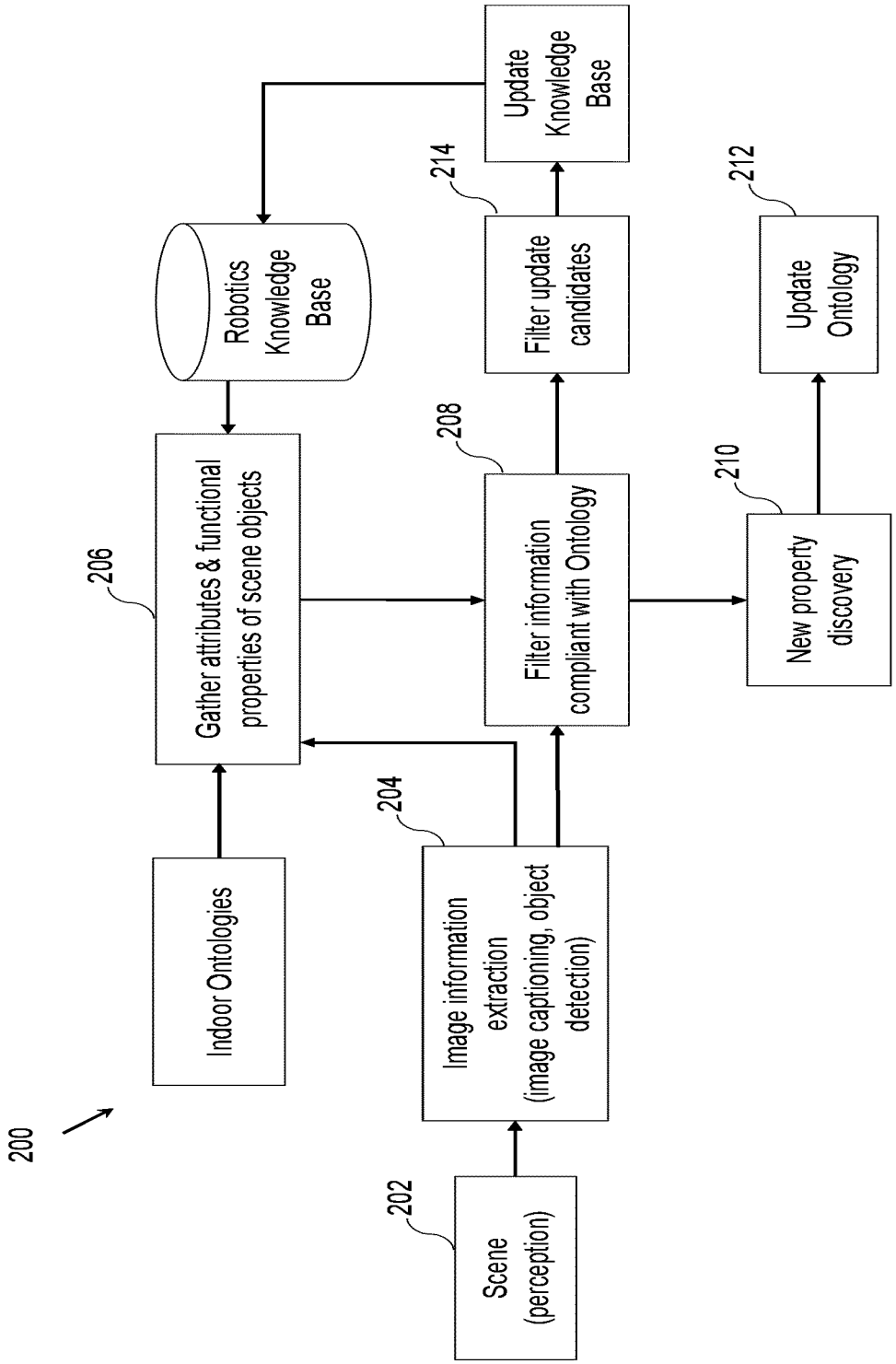
200
Scene (perception)
202
Image information extraction (image captioning, object detection)
204
Indoor Ontologies
Gather attributes & functional properties of scene objects
206
Filter information compliant with Ontology
208
New property discovery
210
Update Ontology
212
Filter update candidates
214
Robotics Knowledge Base
Update Knowledge Base

FIG. 2

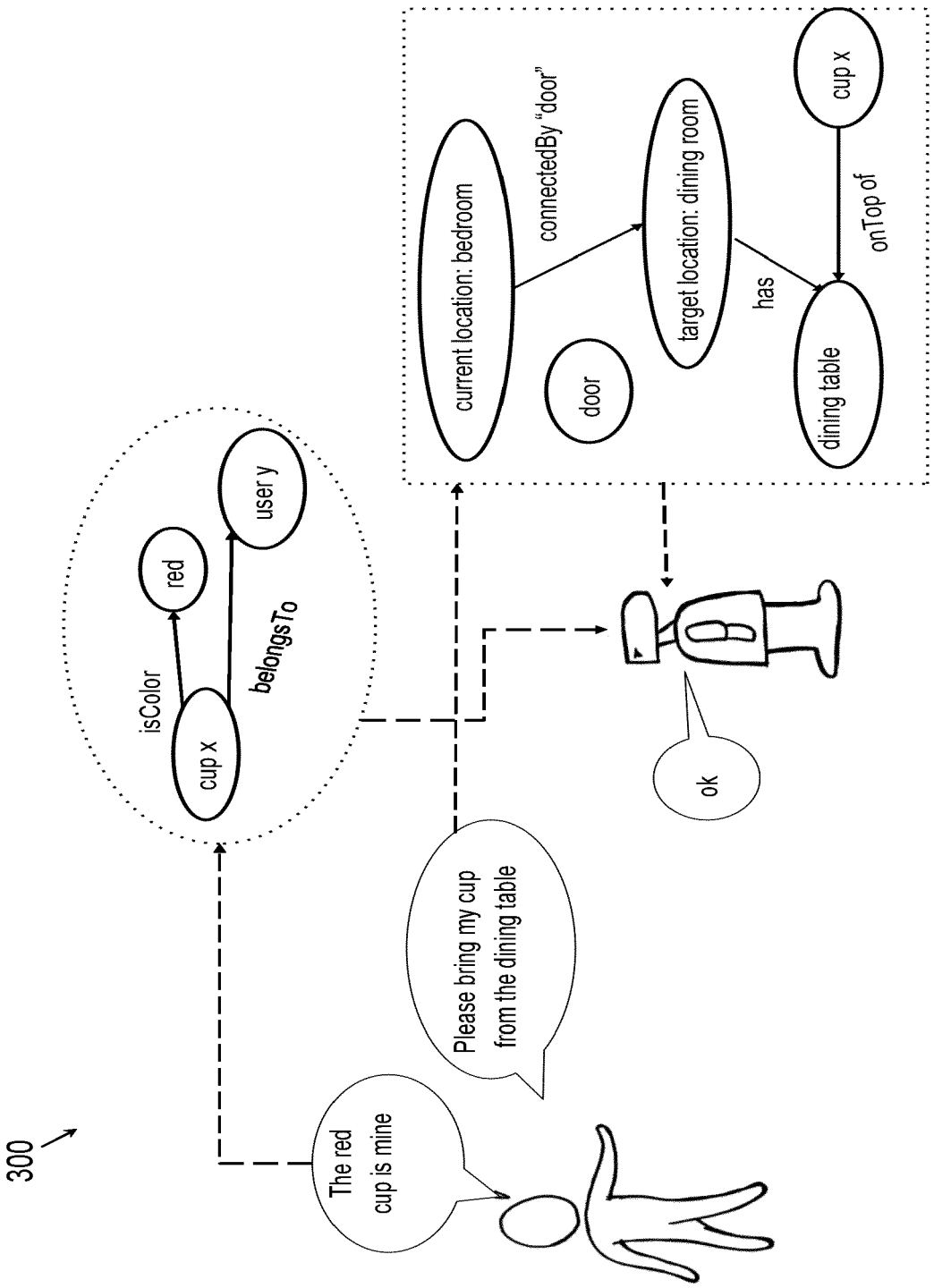
300
The red cup is mine
Please bring my cup from the dining table
ok
cup x
isColor
red
belongsTo
user y
current location: bedroom
connectedBy "door"
door
target location: dining room
has
dining table
onTop of
cup x

FIG. 3

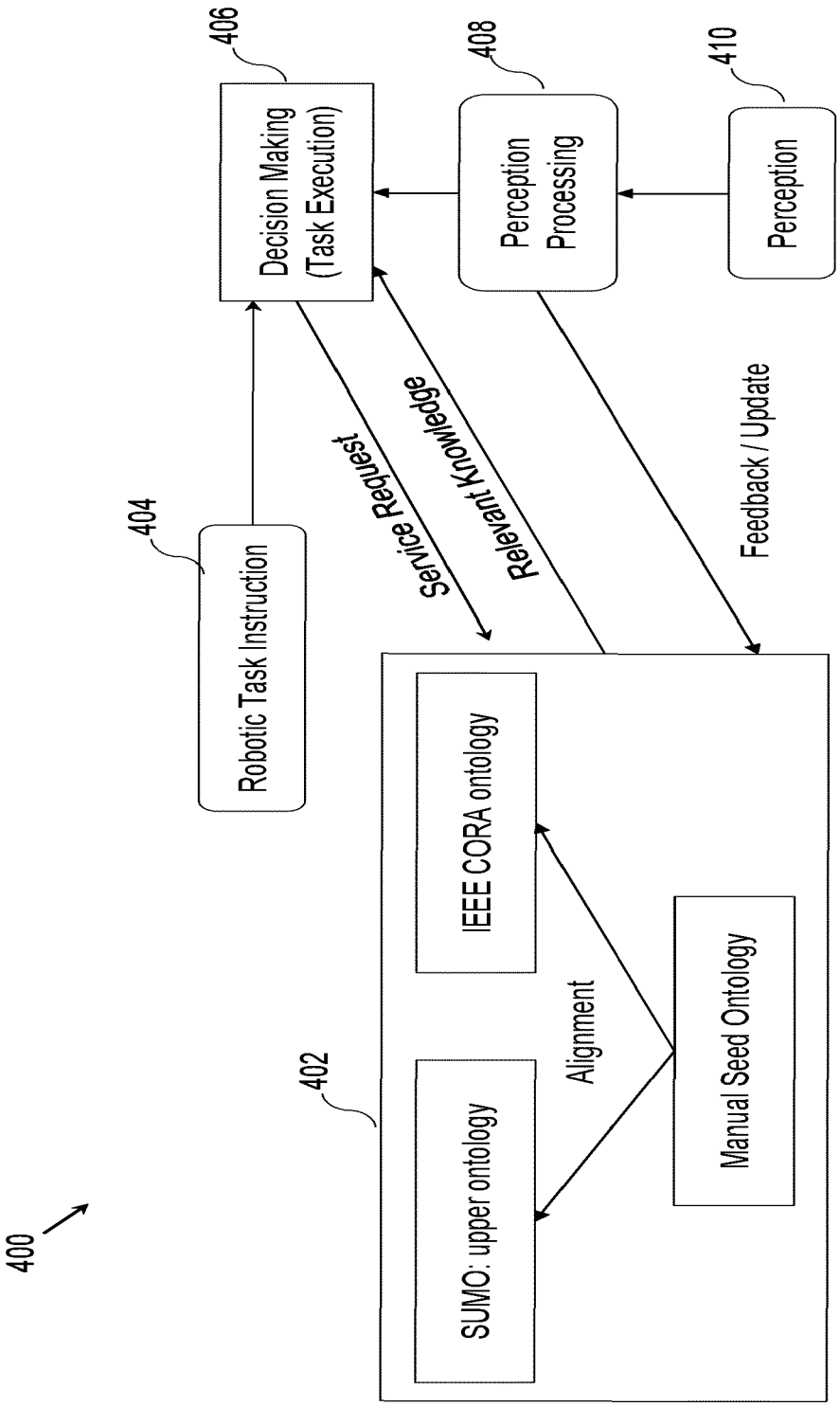
400
402
404
406
408
410
Robotic Task Instruction
Decision Making (Task Execution)
Perception Processing
Perception
Service Request
Relevant Knowledge
Feedback / Update
IEEE CORA ontology
SUMO: upper ontology
Alignment
Manual Seed Ontology

FIG. 4

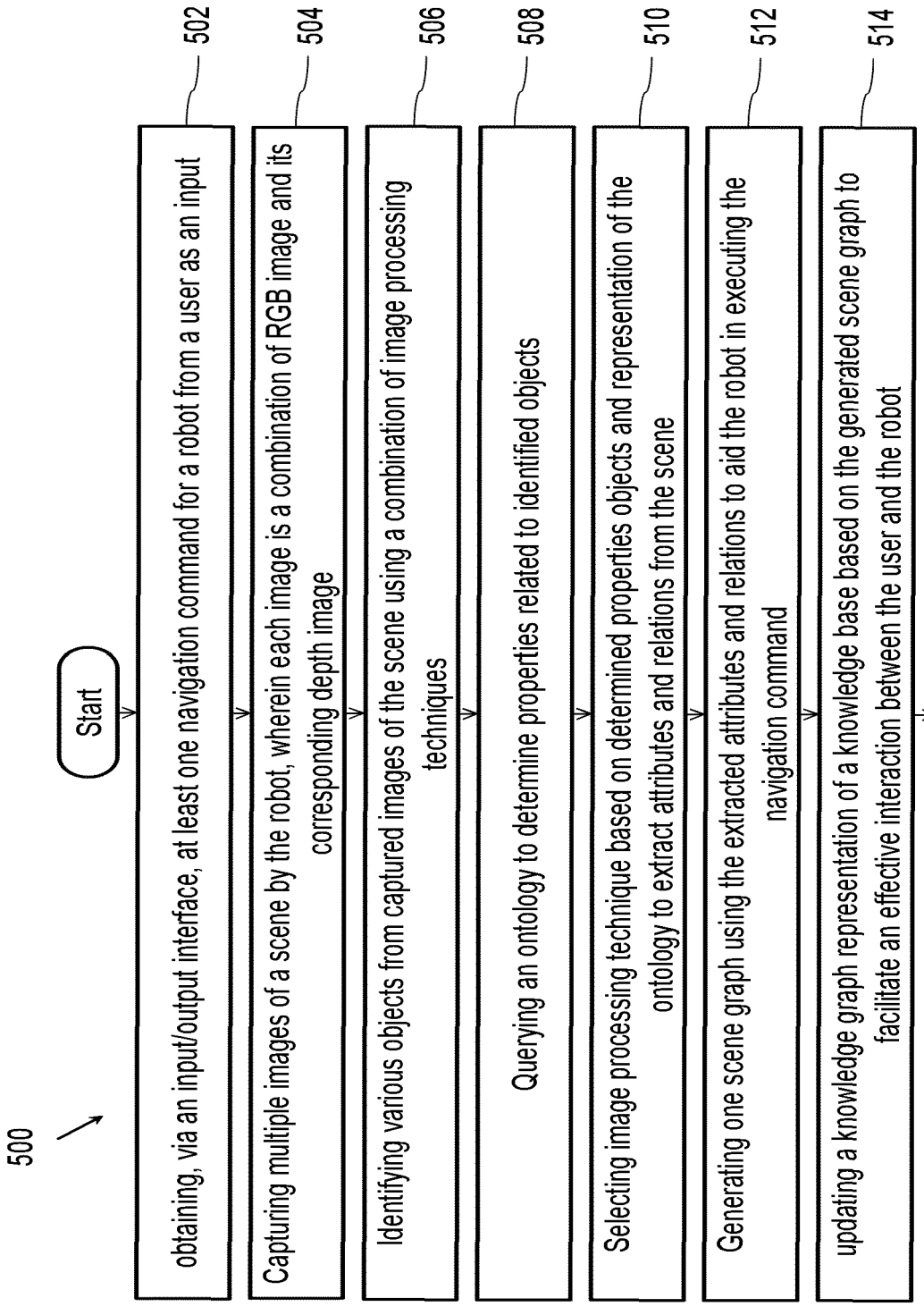
500
Start
obtaining, via an input/output interface, at least one navigation command for a robot from a user as an input
502
Capturing multiple images of a scene by the robot, wherein each image is a combination of RGB image and its corresponding depth image
504
Identifying various objects from captured images of the scene using a combination of image processing techniques
506
Querying an ontology to determine properties related to identified objects
508
Selecting image processing technique based on determined properties objects and representation of the ontology to extract attributes and relations from the scene
510
Generating one scene graph using the extracted attributes and relations to aid the robot in executing the navigation command
512
updating a knowledge graph representation of a knowledge base based on the generated scene graph to facilitate an effective interaction between the user and the robot
514
Stop

FIG. 5

Europäisches Patentamt
European Patent Office
Office européen des brevets

EUROPEAN SEARCH REPORT

Application Number
EP 22 18 6467

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (IPC)
Y	US 2018/043532 A1 (LECTION DAVID B [US] ET AL) 15 February 2018 (2018-02-15) * Paragraphs [0001], [0058], [0061], [0067], [0070], [0071], [0073], [0074]; Figures 4, 8. *	1-15	INV. G05D1/00 G05D1/02 B25J9/16 G01C21/00 G06N5/00 G06V20/00 G10L15/22
Y	US 2019/206400 A1 (CUI RUN [KR] ET AL) 4 July 2019 (2019-07-04) * Paragraphs [0046], [0187], [0241], [0256], [0258]; Figures 19-20. *	1-15	
Y	US 2018/316628 A1 (DEY SOUNAK [IN] ET AL) 1 November 2018 (2018-11-01) * Paragraphs [0025]-[0027], [0039], [0040], [0042]; Figure 1. *	2,5,8, 10,13	
A	CHEN JIANG ET AL: "Understanding Contexts Inside Robot and Human Manipulation Tasks through a Vision-Language Model and Ontology System in a Video Stream", ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 2 March 2020 (2020-03-02), XP081612498, * the whole document *	1-15	

TECHNICAL FIELDS SEARCHED (IPC)
G05D
G06K
G10L
B25J

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
Munich	15 December 2022	Roch, Vincent

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)

ANNEX TO THE EUROPEAN SEARCH REPORT ON EUROPEAN PATENT APPLICATION NO. EP 22 18 6467

This annex lists the patent family members relating to the patent documents cited in the above-mentioned European search report. The members are as contained in the European Patent Office EDP file on
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

15-12-2022

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2018043532 A1	15-02-2018	NONE	
US 2019206400 A1	04-07-2019	NONE	
US 2018316628 A1	01-11-2018	EP 3396607 A1	31-10-2018
		JP 6673958 B2	01-04-2020
		JP 2018190392 A	29-11-2018
		US 2018316628 A1	01-11-2018

EPO FORM P0459

For more details about this annex : see Official Journal of the European Patent Office, No. 12/82

REFERENCES CITED IN THE DESCRIPTION

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Patent documents cited in the description

- IN 202121048303 **[0001]**